複合語の内部情報・外部情報を統合的に利用した訳語対の抽出

吉見　毅彦[†,††] 九津見　毅[†††] 小谷　克則[††]
佐田いち子[†††] 井佐原　均[††]

本稿では，機械翻訳システムの辞書に登録されておらず，かつ (対応付け誤りを含む) 対訳コーパスにおいて出現頻度が低い複合語を対象として，その訳語を抽出する方法を提案する．提案方法は，複合語あるいはその訳語候補の内部から得られる情報と，複合語あるいはその訳語候補の外部から得られる情報とを統合的に利用して訳語対候補に全体スコアを付ける．全体スコアは，複合語あるいはその訳語候補の二種類の内部情報と二種類の外部情報に基づく各スコアの加重和を計算することによって求めるが，各スコアに対する重みを回帰分析によって決定する．読売新聞と The Daily Yomiuri の対訳コーパスを用いた実験では，全体スコアが最も高い訳語対 (のうちの一つ) が正解である割合が 86.36 %，全体スコアの上位二位までに正解が含まれる割合が 95.08 %という結果が得られ，提案手法の有効性が示された．

キーワード:　訳語対抽出，対訳コーパス，低出現頻度語，辞書未登録語，回帰分析

Integrated Use of Internal and External Evidence in the Alignment of Compound Words

Takehiko Yoshimi[†,††] , Takeshi Kutsumi[†††] , Katsunori Kotani[††] ,
Ichiko Sata[†††] and Hitoshi Isahara[††]

This paper proposes a method of extracting English compound words and their Japanese equivalents from a parallel corpus. The aim of our research is to extract compound words which are not listed in a dictionary of an English-to-Japanese MT system and appear infrequently in a parallel corpus. Our method makes its alignment on the basis of two kinds of external evidence provided by the context in which a bilingual pair appears, as well as two kinds of internal evidence within the pair. Each kind of evidence is accompanied by a score, and the aggregate score is computed as a weighted sum of the scores. The appropriate weights are estimated with the logistic regression analysis. An experiment using a parallel corpus of Yomiuri Shimbun and The Daily Yomiuri satisfactorily found that 86.36 % of the extracted bilingual pairs with the highest scores and 95.08 % with the top two scores were judged to be correct.

KeyWords: *Bilingual Lexicon Extraction, Parallel Corpus, Low Frequency Word, Unknown Word, Regression Analysis*

† 龍谷大学, Ryukoku University
†† 情報通信研究機構, National Institute of Information and Communications Technology
††† シャープ (株), SHARP Corporation

1　はじめに

機械翻訳システムの辞書は質，量ともに拡充が進み，最近では200万見出し以上の辞書を持つシステムも実用化されている．ただし，このような大規模辞書にも登録されていない語が現実のテキストに出現することも皆無ではない．辞書がこのように大規模化していることから，辞書に登録されていない語は，コーパスにおいても出現頻度が低い語である可能性が高い．

ところで，文同士が対応付けられた対訳コーパスから訳語対を抽出する研究はこれまでに数多く行なわれ (Eijk 1993; Kupiec 1993; Dekai and Xia 1994; Smadja and McKeown 1996; Ker and Chang 1997; Le, Youbing, Lin, and Yufang 1999)，抽出方法がほぼ確立されたかのように考えられている．しかし，コーパスにおける出現頻度が低い語とその訳語の対を抽出することを目的とした場合，語の出現頻度などの統計情報に基づく方法では抽出が困難であることが指摘されている (辻，芳鐘，影浦 2000)．

以上のような状況を考えると，対訳コーパスからの訳語対抽出においては，機械翻訳システムの辞書に登録されていない，出現頻度の低い語を対象とした方法の開発が重要な課題の一つである．しかしながら，現状では，低出現頻度語を対象とした方法の先行研究としては文献 (辻 2001) などがあるが，検討すべき余地は残されている．すなわち，利用可能な言語情報のうちどのような情報に着目し，それらをどのように組み合わせて利用すれば低出現頻度語の抽出に有効に働くのかを明らかにする必要がある．

本研究では，実用化されている英日機械翻訳システムの辞書に登録されていないと考えられ，かつ対訳コーパス[1]において出現頻度が低い複合語とその訳語との対を抽出する方法を提案する．提案方法は，複合語あるいはその訳語候補の内部の情報と，複合語あるいはその訳語候補の外部の情報とを統合的に利用して訳語対候補にスコアを付け，全体スコアが最も高いものから順に必要なだけ訳語対候補を出力する．全体スコアは，複合語あるいはその訳語候補の内部情報と外部情報に基づく各スコアの加重和を計算することによって求めるが，各スコアに対する重みを回帰分析によって決定する[2]．

本稿では，英日機械翻訳システムの辞書に登録されていないと考えられる複合語とその訳語候補のうち，機械翻訳文コーパス (後述) における出現頻度，それに対応する和文コーパスにおける出現頻度，訳文対における同時出現頻度がすべて1であるものを対象として行なった訳語対抽出実験の結果に基づいて，複合語あるいはその訳語候補の内部情報，外部情報に基づく各条件の有効性と，加重和計算式における重みを回帰分析によって決定する方法の有効性を検証する．

1 本研究で用いたコーパスは，文対応の付いた対訳コーパスであるが，機械処理により対応付けられたものであるため，対応付けの誤りが含まれている可能性がある．

2 回帰分析を自然言語処理で利用した研究としては，重要文抽出への適用例 (Watanabe 1996) などがある

2　訳語対抽出処理の概要

本稿で提案する方法による訳語対抽出処理の概要は次の通りである．

(1)　対訳コーパスのうち英文コーパスを機械翻訳システムで翻訳する．翻訳には「翻訳これ一本 2003」[3] を利用した．

(2)　翻訳結果に原語のまま現れた二単語以上の単語列のうち，大文字で始まる語のみから構成される単語列を対象とする．小文字で始まる語を含む単語列を対象外とする理由は，予備調査の結果，小文字始まり語を含む単語列が辞書に登録されていない原因がつづりの誤りであることと，「kokumin nenkin」のような日本語のローマ字表記であることが多かったからである．大文字始まり語のみから構成される単語列 (複合語) を含む文を抽出し，その集合を機械翻訳文コーパスとする．以下では，大文字始まり語のみから構成される辞書未登録の複合語を単に未登録語と呼ぶ．

(3)　機械翻訳文コーパスとそれに対応する和文コーパスそれぞれに対して形態素解析を行なう．解析には「茶筌」[4] を利用した．

(4)　未登録語に対応する訳語候補を，各機械翻訳文に対応する和文から抽出する．どのような語を訳語候補として抽出するかについては 3.1 節で述べる．

(5)　上記の処理 (2) と (4) で得られた訳語対候補から，機械翻訳文コーパスにおける出現頻度，それに対応する和文コーパスにおける出現頻度，訳文対における同時出現頻度がすべて 1 であるものを抽出する．

(6)　各訳語対候補に対して次の各観点からスコアを付与する．

- 未登録語の構成単語の訳語を単純に合成した訳語と訳語候補との類似性
- 未登録語のローマ字読みと訳語候補の読みとの類似性
- 未登録語の近傍に現れる名詞の集合と訳語候補の近傍に現れる名詞の集合との類似性
- 訳語候補と同一の語が機械翻訳文にも存在するか否か

これらの詳細については，それぞれ 3.2 節，3.3 節，3.4 節，3.5 節で述べる．

(7)　処理 (6) で付与された各スコアを統合して全体でのスコアを決定し，全体スコアが最も高いものから順に必要なだけ訳語対候補を出力する．スコアの統合方法については，3.6 節で述べる．

3 http://www.sharp.co.jp/ej/
4 http://chasen.aist-nara.ac.jp/chasen/

3　訳語対抽出に用いる制約条件と優先条件

本節では，2節で概要を述べた訳語対抽出処理で用いる言語情報 (制約条件と優先条件) について説明する．

3.1　品詞指定による訳語候補の絞込み

辞書未登録の複合語は名詞であることが多い．このため，それに対する訳語候補の品詞も名詞であるとする．ただし，名詞すべてを訳語候補とするのではなく，「茶筌」の細分類品詞のうち，原則として，名詞-一般，名詞-サ変接続，名詞-形容動詞語幹，名詞-副詞可能，名詞-ナイ形容詞語幹，名詞-固有名詞を訳語候補とする．また，「茶筌」で未知語とされた語も，原則として，名詞とみなして訳語候補とする．

これらの原則に従わない主な例外は次の二つの場合である．一つ目は，「する」か「できる」が名詞に後接しているとき，全体をサ変動詞とする場合である．二つ目は，半角の括弧のように記号とみなすべきものが「茶筌」の未知語になっている場合である．

名詞あるいは未知語が連続している場合，全体を複合語とみなして一つの訳語候補とする．なお，名詞あるいは未知語の連続には接辞が含まれていてもよい．以下では，複合語を構成する単語の区切りを '/' で表わす．

3.2　未登録語の構成単語の単純合成訳と訳語候補の類似性による優先順位付け

未登録語の訳語すなわち複合語全体としての訳語は，複合語を構成する個々の単語の訳語を単純に合成したものであるとは限らない．しかし，複合語全体としての訳語と，複合語の構成単語の訳語を単純に合成した訳語との間にはある程度の類似性が見られることもある (熊野，平川 1994; 高尾，富士，松井 1996)．そこで，未登録語の訳語候補と，未登録語の構成単語の訳語を単純に合成した訳語との類似性に応じて未登録語と訳語候補の対にスコアを付けることを考える．以下，未登録語の構成単語の訳語の単純な合成を単純合成訳と呼ぶ．

ここでは，訳語候補と単純合成訳の類似性を表わす尺度としてジャッカード係数 (Romesburg 1992) を用いる．ジャッカード係数は，この場合，未登録語の訳語候補と未登録語の単純合成訳の両方に現れる単語の数を，少なくとも一方に現れる単語の数で割った値であると定義できる．すなわち，未登録語 E の訳語候補 J を構成する単語の集合を X とし，未登録語 E の単純合成訳を構成する単語の集合を Y としたとき，未登録語 E と訳語候補 J の対に対する単純合成訳の類似性スコア S_1 は次の式 (1) で求められる．

$$S_1(E,J) = \frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

なお，本稿では複合語における構成単語の出現順序は考慮しない．

実験に用いた対訳コーパスでは，例えば「Disaster/Prevention/Law」という未登録語の訳語候補として「災害/対策/基本/法」が挙げられる．他方，実験に用いた機械翻訳システムで「Disaster」,「Prevention」,「Law」を個別に翻訳するとそれぞれ「災害」,「防止」,「法」という訳語が得られる．このとき，訳語候補を構成する単語の集合と，単純合成訳を構成する単語の集合の積集合に属するものは「災害」と「法」の2語であり，和集合に属するものは「災害」,「対策」,「基本」,「法」,「防止」の5語である．従って,「Disaster/Prevention/Law」と「災害/対策/基本/法」が対訳である可能性に対して，複合語全体としての訳語と単純合成訳の類似性の観点から S_1(Disaster Prevention Law, 災害対策基本法) $= 2/5$ というスコアが与えられる．

3.3　未登録語のローマ字読みと訳語候補の読みとの類似性による優先順位付け

3.2節で述べた条件は，複合語全体としては辞書に登録されていないが，複合語を構成する個々の単語は辞書に登録されている場合に有効である．しかし，地名や人名，組織名などを表わす固有名詞は辞書に登録されていないことも少なくなく，このような場合には有効に働かない．

今回の実験では読売新聞と The Daily Yomiuri をコーパスとして用いたが，このように日本に関する事柄について述べた記事とその対訳記事を多く含むであろうコーパスを処理対象とする場合には地名や人名，組織名も日本のものであることが多い．このような日本に関する固有表現には日本語をローマ字表記した単語が多く含まれる可能性が高い．実際，The Daily Yomiuri コーパスのうち今回の実験対象とした文を機械翻訳システムで翻訳して得られた訳文に含まれる未登録語から100語を単純無作為抽出し，それらが例えば「Tsukuba/Circuit」のように日本語をローマ字表記した単語を含むかどうかを調べたところ，50語にローマ字表記語が含まれていた．

未登録語が日本語をローマ字表記したものである場合，五十音表を用意すれば，比較的容易にかつある程度の精度でその読みを得ることができると考えられる．また，未登録語の訳語候補の読みも「茶筌」で得ることができる．

これらの点に着目して，未登録語に対して得られるローマ字読みと訳語候補の読みを照合し，スコアを付けることにした．読みのスコアには，3.2節と同じく，ここでもジャッカード係数を用いる．すなわち，未登録語を構成する単語のローマ字読みの集合を X とし，訳語候補を構成する単語の読みの集合を Y としたとき，読みの類似性スコア S_2 を式 (1) と同様の式で求める．

訳語候補については，それを構成する全ての単語の読みが得られるが，未登録語については，ローマ字読みが得られる単語とそうでない単語がある．未登録語を構成する単語のローマ字読みが得られない場合，便宜的にその単語そのものを読みとする．例えば未登録語「Tsukuba/Circuit」の場合,「ツクバ」と「Circuit」を読みとする．従って,「Tsukuba/Circuit」の読みの集合と訳語候補「筑波/サーキット」の読みの集合の積集合に属するものは「ツクバ」となり，和集合に属するものは「ツクバ」,「Circuit」,「サーキット」となるため，未登録語「Tsukuba/Circuit」

と訳語候補「筑波/サーキット」に対する読みの類似性スコア S_2(Tsukuba Circuit, 筑波サーキット) は 1/3 となる.

実験に用いたコーパスでは，未登録語「Tsukuba/Circuit」の訳語候補として,「筑波/サーキット」の他に,「レース/用/バイク」,「パーツ/販売」が挙がってくるが,「レース/用/バイク」と「パーツ/販売」の場合は共に読みの類似性スコアが 0 となるので，未登録語「Tsukuba Circuit」に対する訳語としては「筑波/サーキット」が優先される.

3.4 未登録語の近傍名詞集合と訳語候補の近傍名詞集合との類似性による優先順位付け

訳語候補が未登録語の正しい訳語である場合，和文において訳語候補の前後に現れる語の集合と，機械翻訳文において未登録語の前後に現れる語の集合とは類似性が高いと考えられる (Fung 1998; 梶，相薗 2001)．そこで，未登録語の近傍に現れる名詞の集合と，訳語候補の近傍に現れる名詞の集合との類似性を訳語としての確からしさとして考慮する.

ある語の近傍に現れる語を近傍名詞と呼び，近傍名詞になる可能性がある名詞を近傍名詞候補と呼ぶ[5]．近傍名詞候補には，未登録語の訳語候補 (3.1 節参照) の他に,「茶筌」品詞の名詞-数も含める．これは，例えば「九十二/年」のような数表現は，訳語候補としては適切ではないが，未登録語の訳語候補の中から正しいものを選び出すのには有効な情報を提供すると考えられるからである.

ある語の近傍の範囲は，その語が現れる文に含まれる近傍名詞候補の総数に比例するものとする．今回の実験では，未登録語が現れる機械翻訳文に含まれる近傍名詞候補の総数を N としたとき，未登録語の前方の近傍名詞候補を最大 $N/4$ 語まで，後方の近傍名詞候補を最大 $N/4$ 語まで近傍名詞として集合に加えた．なお，近傍名詞集合において，近傍名詞の出現位置が未登録語の前方か後方かは区別しない．また，近傍名詞候補数の上限値の小数点以下は切り捨てる．訳語候補についての近傍名詞集合についても和文において同様に求める．複合語の語数の計測では，個々の単語に分解せず，複合語全体で一語と数える.

未登録語の近傍名詞集合と訳語候補の近傍名詞集合との類似性スコアもジャッカード係数で表わす．すなわち，未登録語 E の近傍名詞集合を X とし，訳語候補 J の近傍名詞集合を Y としたとき，近傍名詞集合の類似性スコア S_3 を式 (1) と同様の式で求める.

例えば，次の和文 (H1) と機械翻訳文 (M1) から成る組では，未登録語「Maastrich/Treaty」の訳語候補は,「マーストリヒト/条約」と「弾み」である.

(E1) With France, Germany provided a powerful new impetus to European integration, terminating in the Maastrich Treaty of 1992.

(H1) フランスとともにドイツは，九二年のマーストリヒト条約として結実する欧州統

5 近傍名詞候補のうちどれを近傍名詞にするかは，どのくらいの距離を近傍とみなすかによる.

合に新たな，そして強力な弾みを与えた．

(M1) フランスに関して，1992 年の Maastrich Treaty で終了して，ドイツは，欧州統合に強力な新しい刺激をした．

機械翻訳文 (M1) には 7 語の近傍名詞候補が現れる[6]ので，「Maastrich/Treaty」の前後それぞれ 1 語ずつ「1992/年」と「ドイツ」が「Maastrich/Treaty」の近傍名詞となる．他方，和文 (H1) には 8 語の近傍名詞候補が現れる[7]ので，「マーストリヒト/条約」の前方の 2 語「ドイツ」と「九二/年」，および後方の 2 語「欧州統合」と「新た」が「マーストリヒト/条約」の近傍名詞となる．訳語候補「弾み」の場合は，後方に近傍名詞候補が存在しないので，前方の 2 語「新た」と「強力」だけが近傍名詞となる．表 1 に近傍名詞をまとめて示す．

表 1　近傍名詞集合の例

未登録語，訳語候補	近傍名詞集合
Maastrich/Treaty	1992/年，ドイツ
マーストリヒト/条約	ドイツ，九二/年，欧州統合，新た
弾み	強力，新た

未登録語「Maastrich/Treaty」の近傍名詞集合と訳語候補「マーストリヒト/条約」の近傍名詞集合との積集合に属するものは「ドイツ」の 1 語となり，和集合に属するものは，「1992/年」，「ドイツ」，「九二/年」，「欧州統合」，「新た」の 5 語となるので，S_3(Maastrich Treaty, マーストリヒト条約) $= 1/5$ という近傍名詞集合の類似性スコアが与えられる．他方，「Maastrich/Treaty」と訳語候補「弾み」の場合は，両者の近傍名詞集合の積集合に属する単語は存在しないので，近傍名詞集合の類似性スコア S_3(Maastrich Treaty, 弾み) は 0 となる．従って，近傍名詞集合のスコアの観点からは，未登録語「Maastrich/Treaty」の訳語として「マーストリヒト/条約」が「弾み」よりも優先される．

3.5　訳語候補と同一語の存在/非存在による優先順位付け

和文に現れている訳語候補が機械翻訳文にも現れている場合，それらは対応関係にある可能性が高く，訳語候補と未登録語が対応関係にある可能性は低いのではないかと考えられる[8]．例えば，次の和文 (H2) と機械翻訳文 (M2) から成る組では，未登録語「Foodstuff/Sanitation/Law」の訳語候補として，「食品/衛生/法」と「施行/規則」が挙げられる．このうち後者は，機械翻訳文 (M2) に「規則」という単語が存在するため，「Foodstuff/Sanitation/Law」よりも「規則」に

6 機械翻訳文 (M1) に現れる近傍名詞候補は，「フランス」，「1992/年」，「Maastrich/Treaty」，「ドイツ」，「欧州統合」，「強力」，「刺激」である．

7 和文 (H1) に現れる近傍名詞候補は，「フランス」，「ドイツ」，「九二/年」，「マーストリヒト/条約」，「欧州統合」，「新た」，「強力」，「弾み」である．

8 同様の考え方が文献 (石本，長尾 1994) に示されている．

対応する可能性が高いと考えるのが自然であろう．

(E2) First, they plan to coordinate views with their counterparts at the Agriculture, Forestry and Fisheries Ministry and then revise regulations of the Foodstuff Sanitation Law by as early as this fall.

(H2) 同省は、農水省と調整し、この秋にも食品衛生法の施行規則などを改正し、来年四月一日の施行を目指す。

(M2) 最初に、それらは、農林水産省でそれらの相対物を持つビューを統合し、その後、Foodstuff Sanitation Law の規則をこの秋と同じくらい早く改正するつもりである。

このため，機械翻訳文に同じ語が現れる訳語候補と未登録語の対には，機械翻訳文に同じ語が現れない訳語候補と未登録語の対に与えるスコアよりも低いスコアを与える．なお，訳語候補と機械翻訳文に現れる語との照合は，複合語単位ではなく単語単位で行なう．すなわち，訳語候補と機械翻訳文に現れる語の両方あるいは一方が複合語である場合，それらを構成する単語のうち少なくとも一つが一致すれば，両者は対応関係にあるとみなす．実験では，訳語候補を構成する単語と同じ単語が機械翻訳文に現れる場合，同一語の存在/非存在に関するスコア S_4 を 0 とし，現れない場合 0.5 とした．

$$S_4 = \begin{cases} 0 & \text{訳語候補の構成単語と同じ単語が機械翻訳文に現れる場合} \\ 0.5 & \text{現れない場合} \end{cases}$$

3.6 総合的評価

提案方法では，次の式 (2) のような総合評価式に基づいて，未登録語と訳語候補の内部情報 (未登録語の構成単語の単純合成訳と訳語候補との類似性，未登録語のローマ字読みと訳語候補の読みとの類似性) と未登録語と訳語候補の外部情報 (未登録語の近傍名詞集合と訳語候補の近傍名詞集合との類似性，訳語候補と同一語の存在/非存在) を組み合わせた評価を行ない，全体スコア S が最も高い訳語対から順に出力する．

$$S = C + \sum_{i=1}^{4} W_i \times S_i \tag{2}$$

C は定数であり，重み W_i は各観点からのスコア S_i の相対的重要度を表わす．

訳語対内外の情報を併用するという考え方は，文献 (梶，相薗 2001) で示唆されているが，提案の段階に留まっており実験結果などは示されていない．

C や W_i の決定方法として，直感的にあるいは予備実験を通じて経験的に決定する方法 (以下，経験的な決定法と呼ぶ) と，回帰分析によって決定する方法が考えられる．本稿では，両者の場合について訳語対抽出の正解率を比較する．

4　実験と考察

4.1　実験方法

実験には，内山ら (内山，井佐原 2003) によって文対応付けが行なわれた読売新聞と The Daily Yomiuri の対訳コーパスのうち 1989 年から 1996 年 7 月中旬までの記事で構成される部分を用いた．さらに，内山らの文対応スコアの上位 10 %の訳文対に対象を限定した．このコーパスに対して訳語対抽出処理を行ない，各未登録語ごとに全体スコアが高いものから順に訳語候補を出力した．

得られた訳語対データから標本抽出を行ない，標本中の各未登録語とその訳語候補の対に対して次のような「正解」か「不正解」の評価値を与えた．この評価値は，抽出された訳語対を辞書に登録する際の作業量の観点に立ったものである．

正解　訳語の追加や削除，置換を行なわなくても，そのまま辞書に登録できる．
例：Comprehensive/Security/Board $\Longleftrightarrow$ 総合/安全/保障/審議/会

不正解　辞書に登録するためには，訳語の追加や削除，置換が必要である．
例：Liquor/Tax/Law $\Longleftrightarrow$ 逆手

なお，上記の評価を行なう際，次のような場合は対象外とした．この措置は，訳語対抽出に用いた各条件の有効性と統合方法の有効性の検証に重点を置きたいことなどによる．

- 未登録語の抽出が不適切である (未登録語が一つの名詞句を構成していない) 場合．原因は 2 節で述べた処理 (2) が失敗したことにある．
  例：Kita/Ward/Tuesday $\Longleftrightarrow$ 大阪/市/北/区
- 正解の一部分しか訳語候補になっていない場合．処理 (4) の失敗によるものである．なお，この場合は，訳語の追加を行なえば辞書に登録できる．例えば次の例では「法」を追加すればよい．例：Administrative/Procedures/Law $\Longleftrightarrow$ 行政/手続
- 訳語候補に正解が含まれていない場合．これは，機械翻訳文コーパスにおける出現頻度，それに対応する和文コーパスにおける出現頻度，訳文対における同時出現頻度がすべて 1 であるものを対象としているため，処理 (5) で訳語候補から除外されることによる．また，元々，未登録語を含む機械翻訳文に対応する和文に正解が含まれていないことにもよる．
- 未登録語に対する訳語候補が一つしかない場合．訳語候補が正解であっても対象外とした．

総合評価式 (2) における定数 C と重み W_i としては，それぞれ表 2 に示す値を用いた．

経験的な決定法による値は，予備実験で得られた訳語対データから未登録語を 100 語無作為抽出し，各未登録語とその全訳語候補との対に与えられた各条件によるスコアを観察した経験から，訳語対抽出に用いる各条件の信頼性が次の順で高くなっていくと判断したことによるも

表 2　重みの値

	経験的決定法	回帰分析
定数 C	0	−4.58
単純合成訳 W_1	2	20.75
ローマ字読み W_2	3	15.04
近傍名詞集合 W_3	1	3.58
同一語 W_4	2	2.81

のである．

(1)　未登録語の近傍名詞集合と訳語候補の近傍名詞集合との類似性 (S_3)

(2)　未登録語の構成単語の単純合成訳と訳語候補との類似性 (S_1) と，訳語候補と同一語の存在/非存在 (S_4)

(3)　未登録語のローマ字読みと訳語候補の読みとの類似性 (S_2)

回帰分析による方法で決定した値は，個々の条件に基づくスコアを説明変数とし，正解か不正解か (1 か 0 か) を目的変数としてロジスティック回帰分析を行なって求めたものである．訓練データの規模は 1734 件であり，このうち正解が 148 件，不正解が 1586 件である．

表 2 で経験的な決定法による重みの値と回帰分析による方法で決定した重みの値を比較すると，次のような違いがある．

- 経験的な決定法の場合，未登録語と訳語候補の内部情報に基づくスコアに対する重みと未登録語と訳語候補の外部情報に基づくスコアに対する重みの間の差は小さいが，回帰分析による方法の場合は両者の差が比較的大きい．
- 経験的な決定法では各条件の信頼性が $S_3 < S_1 = S_4 < S_2$ のように高くなっていくと考えたが，回帰分析による方法では $S_4 < S_3 < S_2 < S_1$ の順で高くなっていくとみなされている．

4.2　実験結果

抽出された標本は，264 語の未登録語と 1086 語の訳語候補から成る．すなわち，未登録語に対する訳語候補数は平均で 4.11 語 (1086/264) であった．

経験的な決定法による重みを用いた場合と回帰分析による方法で決定した重みを用いた場合のそれぞれの評価結果を表 3 に示す．表 3 を見ると，単独一位での正解率，同点一位を含めた場合の正解率，上位二位までての正解率のいずれにおいても，回帰分析による方法のほうが経験的な決定法よりも高い正解率が得られている．

経験的な決定法と回帰分析による方法とで，正解が出力された順位がどのように変化したかを表わす分布を表 4 に示す．表 4 によれば，経験的な決定法より回帰分析による方法のほうが下がったものが 5 語である．逆に回帰分析による方法のほうが順位が上がったものは 15 語あり，

表 3　訳語対抽出の正解率

	単独一位	第一位 (同点一位を含む)	上位二位
経験的決定法	74.24 %(196/264)	83.71 %(221/264)	92.80 %(245/264)
回帰分析	77.65 %(205/264)	86.36 %(228/264)	95.08 %(251/264)

その内訳は，同点一位から単独一位へ上がったものが 2 語，二位から単独一位へ上がったものが 7 語，三位以下から単独一位へ上がったものが 5 語，三位以下から二位へ上がったものが 1 語である．

表 4　正解が出力された順位の変動

経験的決定法 \ 回帰分析	単独一位	同点一位	二位	三位以下	合計
単独一位	191	0	5	0	196
同点一位	2	23	0	0	25
二位	7	0	17	0	24
三位以下	5	0	1	13	19
合計	205	23	23	13	264

4.3　失敗原因の分析

正解が出力された順位が第二位以下であったものについて，その原因ごとに分類した結果を表 5 に示す．表 5 を見ると，未登録語のローマ字読みと訳語候補の読みとの類似性による優先順位付けの誤りによるもの (ローマ字読み) と訳語候補と同一語の存在/非存在による優先順位付けの誤りによるもの (同一語) の件数が他の原因に比べて多い．この二つの原因について分析する．なお，複合的原因は複数の原因が絡んでいると考えられるものである．

表 5　失敗原因の分類

原因	経験的決定法	回帰分析
単純合成訳	2	6
ローマ字読み	12	11
近傍名詞集合	7	7
同一語	19	9
複合的原因	3	3
合計	43	36

未登録語のローマ字読みと訳語候補の読みとの類似性による優先順位付けの誤りによるものは，次のように細分できる．

- 表記からローマ字読みを得る処理の不備によるもの．実装した処理では，未登録語「Nihon/Shimbun/Kyokai」と正解「日本/新聞/協会」の対応関係が認識できなかった．この原因は，両唇音の直前の閉鎖音は両唇音に変わる音韻規則を考慮していなかったた

め,「shimbun」の読みを得ることができなかったことにある．また,「キョウ」を「kyo」と表記する書記規則を考慮していなかったため,「kyokai」の読みが「キョウカイ」ではなく「キョカイ」になってしまったことにも原因がある．

- 読みの曖昧さによるもの.「Nihon/Shimbun/Kyokai」と「日本/新聞/協会」の対応関係が認識できなかったもう一つの原因は,「日本」の読みが「ニホン」ではなく「ニッポン」になっていたことにある．この問題は,「茶筌」の設定が読みの第一候補だけを得るようになっていたために生じたものであり，全候補を得る設定にすれば解決可能である．
- 「茶筌」辞書の未登録語によるもの.「茶筌」の辞書に「熱川」という固有名詞が登録されていなかったために,「Atagawa」と「熱/川」の対応関係が認識できなかった．

訳語候補と同一語の存在/非存在による優先順位付けの誤りによるものは，機械翻訳文中の訳語候補 (の構成要素) が和文にも現れているにもかかわらず，その訳語候補が正解である場合である．例えば，次の機械翻訳文 (M3) に現れる未登録語「Price/Control/Ordinance」の正解訳語である「物価/統制/令」は，機械翻訳文 (M3) に現れる「物価/上昇」と対応していると誤認識されてしまう．

(E3)　The Price Control Ordinance was imposed in March 1946 to curb price hikes and ease the supply and demand of products.

(H3)　物価統制令は一九四六年三月、物価暴騰を抑え、物資需給の円滑化を目的にポツダム命令として公布。

(M3)　Price Control Ordinance は、物価上昇を抑制し、そして、製品の需要と供給を緩和するために、1946 年 3 月に課された。

このような誤りを防ぐには,「物価/上昇」が「物価/暴騰」に対応していることを認識する必要がある．

4.4　条件間の独立性

複数の条件を組み合わせて訳語対抽出を行なう方法では，条件間の独立性が高いほうが望ましい．そこで，各条件間の独立性を調べるために，各素性間でのスピアマンの順位相関係数 (Siegel 1983) を求めた．その結果を表 6 に示す．

表 6　各条件間でのスピアマンの順位相関係数

	単純合成訳	ローマ字読み	近傍名詞集合	同一語
単純合成訳	−	0.057	0.048	0.039
ローマ字読み	−	−	0.099	0.133
近傍名詞集合	−	−	−	−0.034
同一語	−	−	−	−

表6によれば，どの条件の間でも相関係数の値は小さく，独立性が高いことが分かる．

4.5 各条件の有効性

本節では，各条件が正解率の向上にどの程度寄与しているかを調べる．個々の条件を課さない場合の重みの値は，各条件を除いた状態で1734件の訓練データに対してロジスティック回帰分析を行なうことによって求めた．各条件を課さない場合の正解率を表7に示す．括弧内の数字は正解数である．

表 7 各条件の有効性

	単独一位	第一位 (同点一位を含む)	上位二位
全条件	77.65 %(205)	86.36 %(228)	95.08 %(251)
単純合成訳なし	63.64 %(168)	81.44 %(215)	90.15 %(238)
ローマ字読みなし	58.33 %(154)	79.92 %(211)	90.90 %(240)
近傍名詞集合なし	74.24 %(196)	91.67 %(242)	95.45 %(252)
同一語なし	76.89 %(203)	86.74 %(229)	94.70 %(250)

未登録語の構成単語の単純合成訳と訳語候補との類似性に関する条件を課さない場合と，未登録語のローマ字読みと訳語候補の読みとの類似性に関する条件を課さない場合は，全ての条件を課した場合に比べて，単独一位での正解率，同点一位を含めた場合の正解率，上位二位までての正解率のいずれもが低くなっている．特に単独一位での正解率の低下が大きい．従って，これら二つの条件は正解率の向上に寄与していると言える．

他方，未登録語の近傍名詞集合と訳語候補の近傍名詞集合との類似性に関する条件を課さない場合と，訳語候補と同一語の存在/非存在に関する条件を課さない場合は，全ての条件を課した場合に比べて，同点一位を含めた場合の正解率は高くなっており，また，単独一位での正解率と上位二位までての正解率も若干低くなっている程度である．従って，これら二つの条件は正解率の向上に寄与していないと言える．

正解率の向上に寄与している二つの条件は複合語あるいはその訳語候補の内部情報に関するものであり，寄与していない二つの条件は外部情報に関するものである．訳語対抽出の正解率向上に有効に働く外部情報を探っていくことが今後の課題である．

5 おわりに

本稿では，実用化されている機械翻訳システムの辞書に登録されておらず，かつ，(対応付け誤りを含む) 対訳コーパスにおいて出現頻度が低い複合語を対象として，その訳語を抽出する方法を示した．提案方法では，複合語あるいはその訳語候補の内部の情報とそれらの外部の情報を統合的に利用して訳語対候補に全体スコアを付ける．全体スコアは四種類の情報に基づ

く各スコアの加重和を計算することによって求めたが，各スコアに対する重みをロジスティック回帰分析によって決定する方法を採った．読売新聞と The Daily Yomiuri の対訳コーパスを用いた実験では，加重和による総合評価式において各スコアに対する重みをロジスティック回帰分析により決定した場合，全体スコアが最も高い訳語対 (のうちの一つ) が正解である割合が 86.36 %，上位二位までに正解が含まれる割合が 95.08 %という結果が得られた．この結果は，直感的にあるいは予備実験を通じて経験的に決定する方法による結果を上回るものである．

参考文献

Dekai, W. and Xia, X. (1994). "Learning an English-Chinese Lexicon from a Parallel Corpus." In *Proceedings of the Annual Conference of Association for Machine Translation of America*, pp. 206–213.

Eijk, P. (1993). "Automating the Acquisition of Bilingual Terminology." In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 113–119.

Fung, P. (1998). "Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora." *Lecture Notes in Artificial Intelligence*, **1529**, pp. 1–17.

石本浩之，長尾眞 (1994). "対訳文章を利用した専門用語対訳辞書の自動作成—訳語対応における両立不可能性を考慮した手法について—." 研究報告 NL102-11, 情報処理学会.

梶博行，相薗敏子 (2001). "共起語集合の類似度に基づく対訳コーパスからの対訳語抽出." 情報処理学会論文誌, **42** (9), pp. 2248–2258.

Ker, S. and Chang, J. (1997). "A Class-based Approach to Word Alignment." *Computational Linguistics*, **23** (2), pp. 312–343.

熊野明，平川秀樹 (1994). "対訳文書からの機械翻訳専門用語辞書作成." 情報処理学会論文誌, **35** (11), pp. 2283–2290.

Kupiec, J. (1993). "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora." In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pp. 17–22.

Le, S., Youbing, J., Lin, D., and Yufang, S. (1999). "Word Alignment of English-Chinese Bilingual Corpus Based on Chunks." In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 110–116.

Romesburg, H. C. (1992). 実例クラスター分析. 内田老鶴圃, 東京. 西田英郎，佐藤嗣二 訳.

Siegel, S. (1983). ノンパラメトリック統計学—行動科学のために—. マグロウヒルブック, 東京. 藤本熙 監訳.

Smadja, F. and McKeown, K. (1996). "Translating Collocations for Bilingual Lexicons: A Statistical Approach." *Computational Linguistics*, **22** (1), pp. 1–38.

高尾哲康，富士秀，松井くにお (1996). "対訳テキストコーパスからの対訳語情報の自動抽出." 研究報告 NL115-8, 情報処理学会.

辻慶太 (2001). "対訳コーパスからの低頻度訳語対の抽出：翻字・頻度情報の統合的利用." 第 49 回日本図書館情報学会研究大会発表要綱, pp. 59–62.

辻慶太，芳鐘冬樹，影浦峡 (2000). "対訳コーパスにおける低頻度語の性質：訳語対自動抽出に向けた基礎研究." 研究報告 NL138-7, 情報処理学会.

内山将夫，井佐原均 (2003). "日英新聞の記事および文を対応付けるための高信頼性尺度." 自然言語処理, **10** (4), pp. 201–220.

Watanabe, H. (1996). "A Method for Abstracting Newspaper Articles by Using Surface Clues." In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pp. 974–979.

略歴

吉見毅彦： 1987 年電気通信大学大学院計算機科学専攻修士課程修了．1999 年神戸大学大学院自然科学研究科博士課程修了．(財) 計量計画研究所 (非常勤)，シャープ (株) を経て，2003 年より龍谷大学理工学部情報メディア学科勤務．

九津見毅： 1965 年生まれ．1990 年，大阪大学大学院工学研究科修士課程修了 (精密工学—計算機制御)．同年，シャープ株式会社に入社．以来，英日機械翻訳システムの翻訳エンジンプログラムの開発に従事．

小谷克則： 1974 年生まれ．2002 年より情報通信研究機構受託研究員．2004 年，関西外国語大学より英語学博士取得．

佐田いち子： 1984 年北九州大学文学部英文学科卒業．同年シャープ (株) に入社．現在，同社情報通信事業本部情報商品開発センター技術企画室副参事．1985 年より機械翻訳システムの研究開発に従事．

井佐原均： 1978 年京都大学工学部卒業．1980 年同大学院修士課程修了．博士 (工学)．同年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所関西支所知的機能研究室室長．2001 年情報通信研究機構 (旧：通信総合研究所) けいはんな情報通信融合研究センター自然言語グループリーダー．自然言語処理，機械翻訳の研究に従事．言語処理学会，情報処理学会，人工知能学会，日本認知科学会，ACL，各会員．

(2004 年 1 月 30 日 受付)
(2004 年 4 月 21 日 再受付)
(2004 年 5 月 1 日 採録)